Archetypal Analysis for Nominal Observations

Archetypal Analysis for Nominal Observations

Sohan Seth and Manuel J. A. Eugster

Abstract—Archetypal analysis is a popular exploratory tool that explains a set of observations as compositions of few 'pure' patterns. The standard formulation of archetypal analysis addresses this problem for real valued observations by finding the approximate convex hull. Recently, a probabilistic formulation has been suggested which extends this framework to other observation types such as binary and count. In this article we further extend this framework to address the general case of nominal observations which includes, for example, multiple-option questionnaires. We view archetypal analysis in a generative framework: this allows explicit control over choosing a suitable number of archetypes by assigning appropriate prior information, and finding efficient update rules using variational Bayes'. We demonstrate the efficacy of this approach extensively on simulated data, and three real world examples: Austrian guest survey dataset, German credit dataset, and SUN attribute image dataset.

Index Terms—archetypal analysis, nominal observations, variational Bayes', clustering, prototype, simplex visualization

1 Introduction

Archetype is a form of prototype, i.e., representative observation, that is "an ideal example of a type". Similar to *medoids*, archetypes are interpretable since they relate to actual observations, but compared to medoids (or centroids), archetypes are *extreme* in nature rather than *average*. Other observations are seen as *composition* of archetypes rather than *variation* of prototype. A simple example is the RGB color space: red, green, and blue are the archetypal colors. These colors cannot be composed by other colors, but the three "pure" colors can compose all the other colors in the color space, see Figure 1a for an illustration.

Assessing archetypes brings non-trivial understanding through exploration of *pure* objects (e.g., pure emission sources such as stellar populations, nebular emissions, and nuclear activities that the emission of a galaxy is composed of, see [1]), *unique* attributes (e.g., distinctive facial appearances used for face recognition and face verification, see [2]), and *interesting* aspects (e.g., the archetypal "benchwarmer" when describing and profiling basketball players, see [3]), and thus, it has been a popular exploratory data analysis tool.

Following [4], archetypal analysis can be viewed as finding the minimal convex hull of a set of observations *given* the number of vertices, i.e., the number of archetypes. To elaborate, let $\mathbf{X}$ be a (real-valued) data matrix with each column as an observation. Then, standard archetypal analysis is equivalent to solving the following optimization problem:

$$\min_{\mathbf{W},\mathbf{H}} ||\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{H}||_{\mathrm{F}}^2 \tag{1}$$

with the constraint that both $\mathbf{W}$ and $\mathbf{H}$ are column stochastic matrices. F denotes the Frobenious norm. Given n observations, and K archetypes, $\mathbf{W}$ is a $n \times K$ dimensional matrix and $\mathbf{H}$ is a $K \times n$ dimensional matrix. Here, the columns of $\mathbf{Z} = \mathbf{X}\mathbf{W}$ are the inferred archetypes that exist on the convex hull of the observations due to the stochasticity of $\mathbf{W}$ (actually on the boundary of the convex hull as shown by [4]) and for each j-th sample $\mathbf{x}_j$, $\mathbf{Z}\mathbf{h}_j$ is its projection on the convex hull of the archetypes. [4] further simplified this problem to two alternating nonnegative least squares optimizations: express archetypes as convex combinations of the observations, and express observations as convex combinations of the archetypes. This solution has been extensively used over a decade with only minor modifications (e.g., [5], [6], [7], [8]).

Recently, [9] has suggested a probabilistic extension of this framework (Figure 2). Probabilistic archetypal analysis preserves the principle of archetypal analysis by finding the minimal convex hull in the *parameter space*. To elaborate, assume that $\mathbf{x}_j \sim f(\mathbf{x}|\mathbf{p}_j)$. Different data types, such as binary and count, can be accommodated through suitable choice of observation model f. Let $\boldsymbol{\theta}_j$ be the maximum likelihood estimate of $\mathbf{p}_j$, i.e., $\boldsymbol{\theta}_j$ can be seen as the parametric *profile* that best describes the observation $\mathbf{x}_j$ under model f, when one does not impose any shared latent structure among several observations in the parameter space. Then, probabilistic archetypal analysis is equivalent to the following optimization problem:

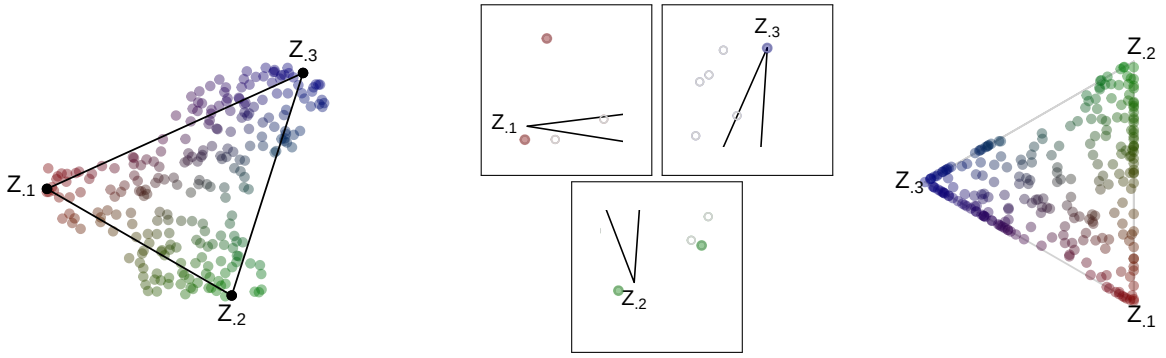
$$\max_{\mathbf{W},\mathbf{H}} \sum_{j=1}^{n} \ln f(\mathbf{x}_j|\Theta\mathbf{W}\mathbf{h}_j) \tag{2}$$

where the $\boldsymbol{\theta}_j$s are columns of Θ. Here, $\mathbf{Z} = \Theta\mathbf{W}$ can be seen as the *archetypal profiles* that exist in the convex hull of Θ (actually on the boundary following the same argument as [4]), and $\mathbf{Z}\mathbf{h}_j$ can be seen as the best approximation of $\boldsymbol{\theta}_j$. It can be easily seen that under (multivariate)

- *S. Seth and M. J. A. Eugster are with Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland E-mail: sohan.seth,manuel.eugster@hiit.fi*

- *The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project.*

- *Implementations of the presented methods and source code to reproduce the presented examples are available at* `http://aalab.github.io`.



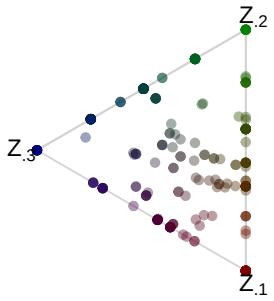
(a) (b) (c)

Obs.	Favorite sports	Marital status	Gender	...
1	Tennis	Single	Female	
2	Golf	Divorced	Male	
3	Soccer	Divorced	Male	
4	Soccer	Divorced	Male	
5	Golf	Married	Female	
6	Tennis	Married	Female	
7	Tennis	Married	Male	
8	Golf	Single	Male	
9	Tennis	Single	Female	
10	Golf	Divorced	Male	
⋮				

(d)

	$Z_{\cdot 1}$	$Z_{\cdot 2}$	$Z_{\cdot 3}$
Favorite sports:			
Golf	0.35	0.00	0.53
Soccer	0.00	1.00	0.00
Tennis	0.65	0.00	0.47
Marital status:			
Single	0.48	0.00	0.00
Divorced	0.52	1.00	0.00
Married	0.00	0.00	1.00
Gender:			
Male	0.37	1.00	0.00
Female	0.63	0.00	1.00
⋮			

(e)



(f)

Fig. 1. Illustration of (top) standard archetypal analysis for real-valued observations and (bottom) the extension we propose for nominal observations. Figure (a) shows a solution with three archetypes $\mathbf{Z}$. Figure (b) shows the generating observations (in color) for each archetype. $\mathbf{Z}_{\cdot 1}$ and $\mathbf{Z}_{\cdot 2}$ have two generating observations, whereas $\mathbf{Z}_{\cdot 3}$ has one generating observation. Figure (c) shows the $\mathbf{H}$ values of the solution in a ternary plot. Observations outside the archetypes solution are projected onto the boundary. Table (d) shows an excerpt from the dataset with nominal features. Table (e) shows the solution with three archetypes. $\mathbf{Z}_{\cdot 3}$, for example, is a married female who likes golf or tennis. A few of the generating observations are visible in (d) with respective colors. Figure (f) shows the $\mathbf{H}$ values; as in (c) the individual samples are expressed as compositions of archetypes.

normal distribution model, i.e., $f \propto \exp(-|\mathbf{x}-\mathbf{p}|^2)$, $\boldsymbol{\theta}_j = \mathbf{x}_j$ and thus, (2) reduces to (1). For other observation models, such as (multivariate) Bernoulli, (multivariate) Poisson and multinomial (*one* nominal variable), (2) can be efficiently solved using majorization-minimization algorithm (for minimizing negative log-likelihood) [9].

1.1 Contribution

Our contributions are, 1. proposing a probabilistic framework to accommodate *multiple nominal variables*, and 2. discussing a principled approach of choosing a suitable number of archetypes by introducing prior information over $\mathbf{W}$ and $\mathbf{H}$.

Contribution 1: Nominal variables appear naturally in response to multiple-choice questions, e.g., in marital status—single, married, divorced, or widowed—or in political view—agree, disagree, or neural. More general examples of nominal variables include bag-of-words representation, e.g., see [10]. We tackle the problem of archetypal analysis when each i-th feature (row of $\mathbf{X}$) is a nominal variable, and thus, each entry $\mathbf{x}_{ij}$ is one (or more) instance(s) of this variable. See Figure 1d for an example. It can be easily verified that the standard formulation in (1) is not applicable here due to the non-Euclidean nature of the nominal variable. Also, (2) can not be applied in a straightforward manner since each nominal variable can have arbitrarily different number of categories.

We achieve this extension with the realization that archetypal analysis with d features is equivalent to d independent archetypal analysis with shared parameters $\mathbf{W}$ and $\mathbf{H}$. To elaborate, the standard archetypal analysis formulation (1), can be written as,

$$||\mathbf{X} - \mathbf{XWH}||_{\mathrm{F}}^2 = \sum_{i=1}^{d} ||\mathbf{X}_{i\cdot} - \mathbf{X}_{i\cdot}\mathbf{WH}||_2^2$$

where $\cdot_{i\cdot}$ denotes i-th row. Therefore, we treat each i-th nominal feature as an independent archetypal analysis problem with multinomial observation model. As a consequence, the Bernoulli and multinomial observation models in [9] are special cases of this framework. Figures 1e-f illustrate archetypal analysis for nominal data following the illustration of standard archetypal analysis for real-valued observations in Figures 1b-c.

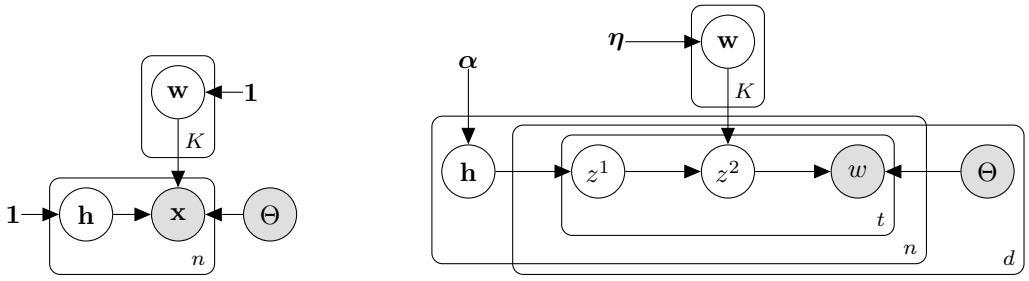


Fig. 2. (a) Plate diagram of probabilistic archetypal analysis as presented in [9]: here $\mathbf{w} \sim \mathrm{Dir}(\mathbf{1}), \mathbf{h} \sim \mathrm{Dir}(\mathbf{1})$ and $\mathbf{x} \sim f(\Theta\mathbf{W}\mathbf{h})$. (b) Plate diagram of probabilistic archetypal analysis for *nominal observations* discussed in this article. We investigate Dirichlet priors over the coefficient vectors: here $\mathbf{w} \sim \mathrm{Dir}(\boldsymbol{\eta})$ and $\mathbf{h} \sim \mathrm{Dir}(\boldsymbol{\alpha})$. Additionally, $z^1 \sim \mathrm{Mult}(\mathbf{h})$, $z^2 \sim \mathrm{Mult}(\mathbf{w}_{z^1})$, and $w^i \sim \mathrm{Mult}(\Theta^i_{\cdot z^2})$. The diagram generalizes two previous cases in [9]: the Bernoulli observation model (f is Bernoulli distribution) is achieved when $v^i = 2$ (v^i being the number of categories in i-th nominal feature) and $t = 1$ for each $i = 1, \ldots, d$ (t being the number of instances of feature i), whereas the multinomial observation model (f is multinomial distribution) is achieved when $d = 1$ (d being the number of features). However, it also allows other observation models such as multiple-choice questions when $t = 1$, but d and v^i can be arbitrary, and multi-view textual representation when d, t and v^i are all arbitrary.

Contribution 2: Probabilistic archetypal analysis assumes that both $\mathbf{w}_k, k \in \{1, \ldots, K\}$ and $\mathbf{h}_j, j \in \{1, \ldots, n\}$ originate from symmetric Dirichlet distribution with concentration parameters n and K respectively, which do not effect the maximum likelihood solution. We impose explicit prior information by varying these coefficients, and study the approximate posterior distribution of $\mathbf{w}_k$ and $\mathbf{h}_j$ using variational Bayes'. We show that this provides a principled approach toward selecting an appropriate number of archetypes, which has previously been done by the popular yet subjective "elbow criterion". To demonstrate the efficacy of this approach, we compare it against the maximum likelihood solution of the same model. We show that while expectation maximization solution works equally well in finding the correct archetypes, it has difficulty in finding suitable number of archetypes. On the other hand, the variational Bayes' solution can effectively select suitable number of archetypes. We show the efficacy of the proposed set-up on multiple real world questionnaire datasets, and discuss the archetypes found.

2 METHOD

Consider that we have a ($d \times n$ dimensional) set of observations $\mathcal{X}$ where each column is one of n observations, and each row is one of d features. Each feature i is a nominal variable with v^i categories $\mathcal{C}^i = \{c^i_1, \ldots, c^i_{v^i}\}$. Thus, each element $\mathcal{X}_{ij}$ is one or more instances of this variable, i.e., $\mathcal{X}_{ij} = \{w^{ij}_1, \ldots, w^{ij}_t, \ldots, w^{ij}_{t^{ij}}\}$ and $w^{ij}_t \in \mathcal{C}^i$. t^{ij} is the total number of instances in $\mathcal{X}_{ij}$. In this paper, we particularly focus on $t^{ij} = 1$. See Figure 1d for an example of $\mathcal{X}$.

Let Ψ^i_{kj} be the number of c^i_k in $\mathcal{X}_{ij}$. Thus $\sum_k \Psi^i_{kj} = t^{ij}$, and each Ψ^i can be seen as a $v^i \times n$ dimensional count matrix. Given this count matrix, the corresponding Θ^i can be estimated as follows:

$$\Theta^i_{kj} = \frac{\Psi^i_{kj}}{\sum_k \Psi^i_{kj}} \forall k = 1, \ldots, v^i; i = 1, \ldots, d; j = 1, \ldots, n.$$

Thus Θ^i can be seen as a column stochastic matrix of dimension $v^i \times n$.

Then, following principles of probabilistic archetypal analysis, we need to solve the following optimization problem:

$$\max_{\mathbf{W},\mathbf{H}} \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{v^i} \Psi^i_{kj} \log(\Theta^i \mathbf{W}\mathbf{H})_{kj} \tag{3}$$

with the constraint that both $\mathbf{W}$ and $\mathbf{H}$ are column stochastic. Notice that under this formulation, the Bernoulli observation model is a special case with $v^i = 2$, and multinomial observation model is a special case with $d = 1$ [9].

This formulation can be also be viewed as follows: for each category w in the i-th subproblem (row) and j-th sample (column), $w \in \mathcal{X}_{ij}$,

1) Choose an archetype, $z^1 \sim \mathrm{Mult}(\mathbf{h}_j)$,
2) Choose a profile, $z^2 \sim \mathrm{Mult}(\mathbf{w}_{z^1})$,
3) Choose a category, $w \sim \mathrm{Mult}(\Theta^i_{\cdot z^2})$.

Here $z^1 \in \{1, \ldots, K\}$ and $z^2 \in \{1, \ldots, n\}$. This ensures that each category $w^{ij}_t = c^i_k$ originates with probability $\mathbf{P}^i_{kj}$ where $\mathbf{P}^i = \Theta^i \mathbf{W}\mathbf{H}$. The archetypal profiles $\{\mathbf{Z}^i = \Theta^i \mathbf{W}, i = 1, \ldots, n\}$ are then found by maximizing the likelihood (3).

We extend the generative framework by adding prior information to $\mathbf{W}$ and $\mathbf{H}$ as follows:

$$\mathbf{h}_j \sim \mathrm{Dir}(\boldsymbol{\alpha}) \forall\, j = 1, \ldots, n$$
$$\mathbf{w}_k \sim \mathrm{Dir}(\boldsymbol{\eta}) \forall\, k = 1, \ldots, K$$

Dirichlet distribution is a natural choice due to its conjugacy to multinomial distribution. Moreover we use

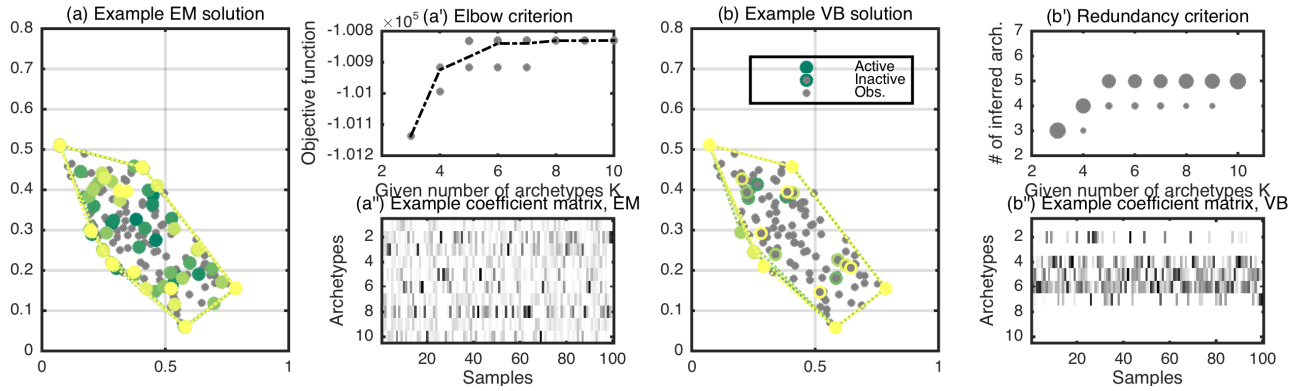


Fig. 3. The figures illustrate the difference between EM and VB approaches. Figures (a) and (b) show multiple EM and VB solutions (shown in different colors) achieved on the same set of observations on a 2 dimensional simplex with $K = 10$ archetypes. Figure (a') shows the variation of objective function with varying number of archetypes K: it stabilizes at $K = 6$. Figures (a") and (b") show the coefficient matrix $\mathbf{H}$ for EM and VB solution. While EM solution assigns weights to all archetypes, VB solution sets certain weights to zero: the archetypes with nonzero factor values are called 'active'. Figure (b') shows the number of active archetypes from different trials for each K. Since the problem is simple the VB solution almost always indicates to the correct number of archetypes.

symmetric prior such that we only have two hyperparameters to control, i.e., α and η.

We infer the approximate posterior distribution using variational Bayes' principle. For archetypal analysis, the variational lower bound is given by,

$$\mathbb{E}_q[\log p(\mathcal{X}, \mathcal{Z}, \mathbf{W}, \mathbf{H}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \Theta)] + \mathbb{H}(q)$$

where

$$q(\mathbf{W}, \mathbf{H}, \mathcal{Z}) = \prod_{k=1}^{K} q(\mathbf{w}_k) \prod_{j=1}^{n} q(\mathbf{h}_j) \prod_{i=1}^{d} \prod_{j=1}^{n} \prod_{t=1}^{t^{ij}} q(z_\varrho^1, z_\varrho^2)$$

where we have used $\varrho := w_t^{ij}$ due to better readability, and $\Theta = \{\Theta^i\}$. We estimate the variational distributions as

$$q(\mathbf{w}_k) \sim \mathrm{Dir}(\boldsymbol{\varsigma}_k),\ q(\mathbf{h}_j) \sim \mathrm{Dir}(\boldsymbol{\beta}_j),\ (z_\varrho^1, z_\varrho^2) \sim \mathrm{Mult}(\boldsymbol{\phi}^{ijl})$$

where $\boldsymbol{\phi}^{ijl}$ are distributions over $\{1, \ldots, K\} \times \{1, \ldots, n\}$, and $l = c(t)$, i.e., the category of the t-th instance.

The variational lower bound can be evaluated as,

$$\mathbb{E}_q \log p(\mathcal{X}, \mathcal{Z}, \mathbf{W}, \mathbf{H}|\boldsymbol{\alpha}, \boldsymbol{\eta}, \Theta) = \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{l=1}^{v^i} \Psi_{lj}^i \times$$

$$\left(\sum_{m=1}^{n} \sum_{k=1}^{K} \boldsymbol{\phi}_{km}^{ijl} \left(\log \Theta_{lm} + \mathbb{E}_q \log \mathbf{W}_{mk} + \mathbb{E}_q \log \mathbf{H}_{kj} \right) \right)$$

$$+ \sum_{m=1}^{n} \sum_{k=1}^{K} (\alpha_m - 1) \mathbb{E}_q[\log \mathbf{W}_{mk}]$$

$$+ \sum_{k=1}^{K} \sum_{j=1}^{n} (\beta_k - 1) \mathbb{E}_q[\log \mathbf{H}_{kj}]$$

$$- \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{l=1}^{v^i} \Psi_{lj}^i \sum_{m} \sum_{k} \boldsymbol{\phi}_{km}^{ijl} \log \boldsymbol{\phi}_{km}^{ijl}$$

$$+ \sum_{k=1}^{K} \mathbb{H}(\mathbf{w}_k) + \sum_{j=1}^{n} \mathbb{H}(\mathbf{h}_j).$$

Maximizing this expression, the corresponding update rules are given by

$$\phi_{km}^{ijl} \propto \exp(\log \Theta_{lm}^i + \mathbb{E}_q \log \mathbf{W}_{mk} + \mathbb{E}_q \log \mathbf{H}_{kj})$$

$$\varsigma_{mk} = \eta + \sum_{i=1}^{d} \sum_{j=1}^{n} \sum_{k=1}^{v^i} \Psi_{lj}^i \phi_{km}^{ijl}$$

$$\beta_{kj} = \alpha + \sum_{i=1}^{d} \sum_{m=1}^{n} \sum_{l=1}^{v^i} \Psi_{lj}^i \phi_{km}^{ijl}.$$

2.1 Hyperparameters

The variational Bayes' solution provides a principled approach for finding suitable number of archetypes. The intuition behind this is as follows: probabilistic archetypal analysis finds the minimal convex hull of a set of parameter values. If one selects more archetypes than actually needed, these archetypes can be placed anywhere inside the convex hull without effecting the final outcome in terms of likelihood value. Also, since they are redundant, the corresponding factor values in $\mathbf{H}$ can be arranged in multiple ways to reach the same solution. Under such circumstances, while the point estimate (maximum likelihood) reaches any of these possible solutions, a sparser and more meaningful solution can be found by utilizing appropriate prior information. Thus, the posterior means of the factor values provide a clue to which archetypes are relevant or 'active' for an observation, and the number of active components can be controlled by the hyperparameters.

Let us first examine the effect of α. Consider the approximate convex hull of profiles $\boldsymbol{\theta}_j$, $j = 1, \ldots, n$ with vertices $\Theta \mathbf{w}_k$, $k = 1, \ldots, K$. The true profiles are

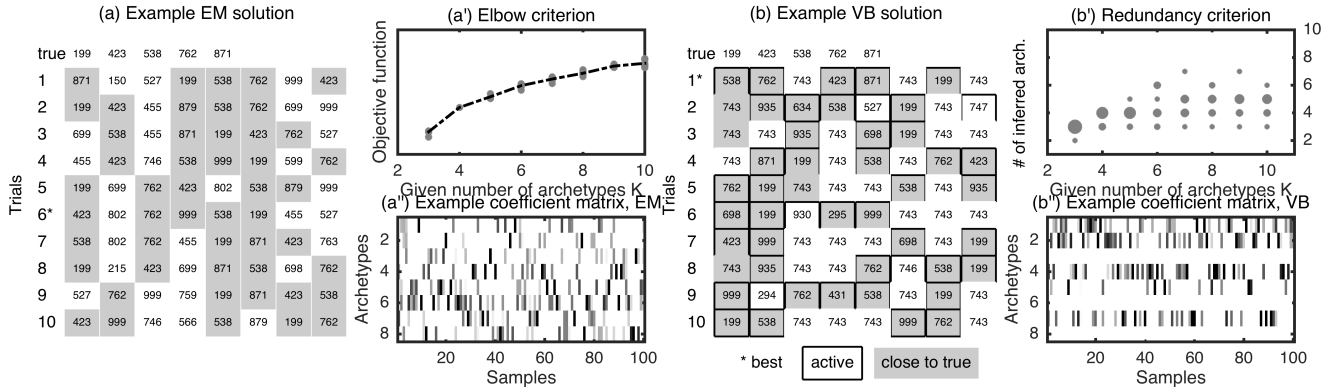


Fig. 4. The figures illustrate the difference between EM and VB approaches. Figures (a) and (b) show multiple EM and VB solutions achieved on the same set of binary observations in 8 dimensions with $K = 8$ archetypes: the archetypes are binarized and then shown in decimal. Figure (a') shows the variation of objective function with varying number of archetypes: this displays a monotonic increase in likelihood value without a clear elbow. Figures (a") and (b") show the coefficient matrix $\mathbf{H}$ for EM and VB solution. While EM solution assigns weights to all archetypes, VB solution sets certain weights to zero: the archetypes with nonzero factor values are called 'active'. Figure (b') shows the number of active archetypes from different trials for each K. Since the problem is difficult the VB solution indicates different number of archetypes but around the right ballpark. If we consider the best objective value then it finds the correct archetypes.

then represented in terms of projections $\mathbf{h}_j$, $j = 1, \ldots, n$. Now, if a profile exists outside the convex hull, then the corresponding projection values are sparse, because this profile is projected on one of the surfaces of the hull, and thus, the factors corresponding to non-contributing vertices of that hyperplane are zero. On the other hand, if a profile is inside the convex hull then there is a possibility that more vertices are contributing to this profile. Therefore, a small value of α encourages profiles to lie outside the convex hull—in other words, it shrinks the convex hull—whereas, a large value of α encourages profiles to lie inside the convex hull—in other words, it inflates the convex hull, and in extreme case it overfits the profiles. Since our objective is to automatically set the factors corresponding to some loadings to zero, i.e., to encourage sparsity, we suggest setting $\alpha < 0.5$. In our experiments, we set this value to 0.3. In practice, however, the loadings are not exactly zero. Therefore, we consider an archetypal profile (k-th) to be 'active' if the maximum projection value related to this archetype over all observations, i.e., $\max_j \mathbf{H}_{kj}$, is greater than 0.15. Also, unless otherwise stated, we always use $K = 20$, i.e., we run the variational Bayes' update rules for 20 archetypes, and select active archetypes as a post-processing step. To show the dependence of inferred number of archetypes on the hyperparameter α, we generate $n = 200$ binary observations in $d = 10$ dimensions with 8 true archetypes, and vary the hyperparameter values in the range $(0.1, 0.2, \ldots, 1)$. We observe that the inferred number of archetypes increases monotonically as the hyperparameter value is increased. We present the result in Figure 5.

Similarly, a large η encourages sharing profiles (non-sparse $\mathbf{w}_k$) to construct archetypal profile whereas small

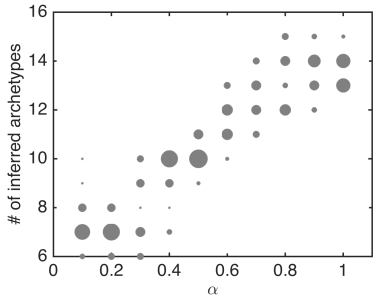


Fig. 5. The figure illustrates the dependence of number of inferred archetypes on hyperparameter α as a "bubble chart", the width of the bubble is the fraction of times the number of inferred archetypes is y given hyperparameter value x. We observe that as hyperparameter is increased, more active archetypes have been found.

η prefers lack of sharing (sparse $\mathbf{w}_k$): therefore, a large η assists the convex hull to shrink. However, more importantly, the function of $\mathbf{w}$ is to construct an archetypal profile from the profiles, and it is usually a very sparse vector–set to a small value to assist sparsity of this vector. We suggest setting $\eta = 0.1$.

2.2 Related methods

Topic modeling: The proposed formulation shares significant resemblance with probabilistic latent semantic analysis [11], and its extension latent Dirichlet allocation [10]. To summarize, PLSA achieves the decomposition $\mathbf{X} = \mathbf{Z}_0\mathbf{H}$ compared to archetypal analysis that achieves the decomposition $\mathbf{X} = \Theta\mathbf{W}\mathbf{H}$, thus in essence the archetypal profiles $\mathbf{Z} = \Theta\mathbf{W}$ inferred by archetypal analysis are data-driven topics. Notice that here we have

only discussed the case when $d = 1$, i.e., the single view case. For $d > 1$, multi-view extension of LDA has also been explored in the literature [12].

The dependence of topics on observations plays a significant role in the interpretability of the topics: while in PLSA or LDA the topics can be abstract concepts, in AA the topics are more 'grounded' (related to original documents through $\mathbf{W}$). Thus, in AA the topics can be explained by the documents that are actually contributing to the topic, whereas in PLSA/LDA the topics might not be (and often not) related to some specific documents. That said, the solution achieved by PLSA/LDA is nonetheless interpretable in its own way (usually by observing the top set of words) and is a widely popular tool for exploring abstract concepts that compose a document.

Finding prototypes: Archetypal analysis offers an alternate, and arguably more interpretable, representation of a set of observations. To elaborate, consider that we have three groups of observations that are well separated from each other, i.e., they form well defined clusters. Then one can choose the centroid of a set of observations (in a cluster) as the representative (or, a typical example) of the corresponding group. Now consider the case when these groups are not well separated. In this situation, although centroids remain a valid representation of the groups for computational purposes, their interpretability degrades since now they are closer to each other than before. Here, it is arguably more reasonable to choose the representatives of the groups (or, typical example of the group) to be more *extreme*. This makes the representatives easier to interpret since now they are further apart from each other. Archetypal analysis finds the latter representation.

The intuitive nature of archetypal analysis and its connection to *topics* and *cluster centers* are illustrated in Figure 7. Here the observations lie on a probability simplex (only first two dimensions are shown). We show the topics inferred by LDA, the archetypal profiles inferred by probabilistic AA, and the cluster centers inferred by k-means. We notice that topics can be abstract concepts whereas archetypal profiles are closely related to the observations. Also, the cluster centers may not lead to meaningful representation when the observations are close to each other. Although these aspects are rather intuitive here, this intuition *seemingly* breaks down for nominal observations since, for example, in case of binary data the observations exist on corners of a hypercube in d-dimensions. However, in the later part of next section, we demonstrate that the intuition of archetypes prevails in this extreme case as well.

3 Simulation

Comparison with EM solution: To illustrate the advantage of the proposed approach, we compare it against the point estimate of $\mathbf{W}$ and $\mathbf{H}$ achieved by maximizing (3) using expectation maximization. For archetypal analysis model, we have

$$\begin{aligned}&\log p(\mathcal{X}, \mathcal{Z}|\mathbf{W}, \mathbf{H}, \Theta)\\ &= \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{t=1}^{t^{ij}} \left(\log \Theta_{\varrho z^2_\varrho} + \log \mathbf{W}_{z^2_\varrho z^1_\varrho} + \log \mathbf{H}_{z^1_\varrho j}\right)\end{aligned}$$

with $\mathcal{Z}_{ij} = \{(z^1_\varrho, z^2_\varrho); t = 1, \ldots, t^{ij}\}$, and $\Theta = \{\Theta^i\}$. Here we have used $\varrho := w^{ij}_t$ for better readability. With slight abuse of notation, $\Theta^i_{\varrho\cdot}$ is the $(k, \cdot)$ entry of Θ^i where $\varrho = c^i_k$. Then, $\mathbf{W}$ and $\mathbf{H}$ can be inferred using expectation-maximization algorithm with the following update rules:

$$\begin{aligned}\mathbf{H}^{t+1}_{kj} &= \sum_{i=1}^{d}\sum_{l=1}^{n}\sum_{m=1}^{v^i} \frac{\mathbf{X}^i_{mj}\Theta^i_{ml}\mathbf{W}_{lk}\mathbf{H}_{kj}}{(\Theta^i\mathbf{W}\mathbf{H})_{mj}},\\ \mathbf{H}^{t+1}_{kj} &= \frac{\mathbf{H}^{t+1}_{kj}}{\sum_k \mathbf{H}^{t+1}_{kj}}\end{aligned}$$

and

$$\begin{aligned}\mathbf{W}^{t+1}_{jk} &= \sum_{i=1}^{d}\sum_{l=1}^{n}\sum_{m=1}^{v^i} \frac{\mathbf{X}^i_{ml}\Theta^i_{mj}\mathbf{H}_{kl}\mathbf{W}_{jk}}{(\Theta^i\mathbf{W}\mathbf{H})_{ml}},\\ \mathbf{W}^{t+1}_{jk} &= \frac{\mathbf{W}^{t+1}_{jk}}{\sum_j \mathbf{W}^{t+1}_{jk}}.\end{aligned}$$

The likelihood of the model increases monotonically as more archetypes are added, since that better approximates the true convex hull. Therefore, to choose an appropriate number of archetypes one usually follows the 'elbow criterion' (terminology usually used while minimizing error rather then maximizing likelihood): find *several* solutions with increasing number of archetypes, and observe if the resulting likelihood value stabilizes. If the problem is simple, it is expected to happen (e.g., Figure 3), whereas if the problem is difficult then a proper plateau may not be observed (e.g., Figure 4). Thus, although quite popular, this approach is not infallible, and also since it is based on visual inspection, it is subject to human error. We demonstrate this aspect on two simulated examples: one where there exists a clear convex hull, and the other where the convex hull is vague.

In the first example, we consider selecting appropriate number of archetypes on 2 dimensional probability simplex. We generate 5 equispaced samples on a circle on the 2 dimensional simplex, which act as archetypal profiles. We generate $n = 100$ samples in the convex hull formed by the archetypes: for each of these resulting probability vector we generate multinomial observations with arbitrary number of trials, generated from a Poisson distribution with rate 1000. The large rate ensures that the empirical distribution is close to the true distribution. Thus, we need to decompose a 3×100 dimensional count matrix. We present the solution achieved by $K = 10$ archetypes in Figure 3. It is observed that VB can effectively find the correct number of archetypes by 'shutting down' redundant archetypes provided we start with sufficiently large number of archetypes.

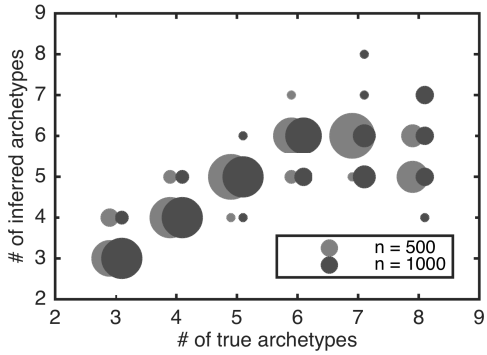


Fig. 6. The figure shows the inferred number of archetypes for varying number of true archetypes found by VB solution for two different sample sizes n as a "bubble chart", the width of the bubble is the fraction of times the number of inferred archetypes is y given true number of archetypes x.

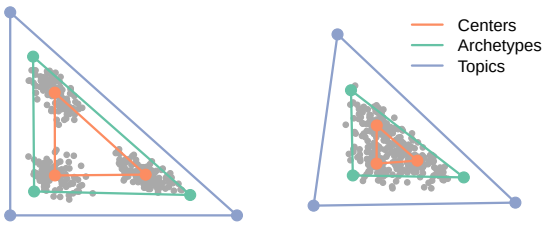


Fig. 7. Comparison among topic modeling, probabilistic archetypal analysis, and clustering. The gray circles are observations, blue circles are topics, green circles are archetypal profiles, and red circles are centroids. We observe that: 1) archetypal profiles offer better interpretability when the observations are close to each other, and 2) topics can be abstract whereas archetypal profiles are closely related to the observations. Refer to Section 2.2 for more information.

In the second example, we consider selecting appropriate archetypes for binary data. Since binary observations exist on corners of a hypercube, there is no explicit convex hull, and thus selecting appropriate number of archetypes is more challenging. We randomly generate 5 binary archetypes in $d = 10$ dimension, and generate $n = 100$ binary observations from probability values within the convex hull of the binary archetypes. We represent the solution achieved by $K = 10$ archetypes in Figure 4. It is observed that the number of 'active' components in VB solution strongly agrees with the correct number of archetypes over multiple trials, whereas EM solution displays a monotonic increase in likelihood without a clear plateau.

To further explore if we can deduce the correct number of archetypes, we randomly generate binary observations in $d = 16$ with T archetypes. For each observation set we find the best archetypal solution (in terms of maximum lower bound) over 10 random initializations with $K = 20$ archetypes. We present the result in Figure 6. We observe that the inferred number of archetypes is highly accurate when T is small. However, when T is large, the inferred number of archetypes drops. This is expected since as the number of true archetypes grows the binary observations gets scattered around the hypercube, and a fixed number of observations do not contain sufficient evidence for detecting the archetypes reliably. In such situation, it is actually preferable that a regularized solution is estimated.

Comparison with standard AA We demonstrate the efficacy of the proposed approach over standard archetypal analysis framework by generating binary observations in $d = 10$ dimensions. We generate $n = 100$ observations from $K = 6$ archetypes, and find archetypal profiles using both the proposed framework and standard archetypal analysis framework. We evaluate both methods in terms of how many inferred archetypes uniquely (no two inferred archetype match the same true archetype) match a true archetype with maximum log-likelihood. Since the dataset does not display a clear 'elbow criterion', we use $K = 6$ for standard formulation, while for the proposed formulation, we use $K = 20$ as mentioned before. We observe that the proposed approach has been able to infer the true archetypes with reasonable accuracy: $4.8/6$ on an average (using $K = 20$ archetypes), compared to $5.4/6$ on an average for standard formulation (using $K = 6$ archetypes). Notice that we have not a priori specified the true number of archetypes for the former, but have done so for the latter.

Comparison with binary clustering: To further explore the difference between archetypal analysis and clustering for nominal observations we compare it against clustering of *binary observations*. We use EM instead of VB to keep the number of prototypes same over all datasets for consistent visualization, and qualitative assessment. We use the tool BernoulliMix[1] [13] for clustering. We keep the number of prototypes the same, 4, for both methods, i.e., four cluster centers and 4 archetypal profiles. We use four datasets, 1) DNA: 342 samples in 12 dimensions, available in the BernoulliMix package, 2) SPECT: 267 samples in 22 dimensions, 3) CONGRESS: 435 samples in 16 dimensions, both congress and spect are available at the UCI machine learning repository [14], and 4) RANDOM: 200 samples in 25 dimensions, each entry randomly generated with probability 0.5. Each former dataset has a clearer cluster structure than the latter, i.e., while DNA has a very clear cluster structure, RANDOM does not have any, and the other two datasets fall in between. We present the solution achieved by the two methods in Figure 8. We observed the following aspects which are in line with our general intuition about archetypes.

First, there is a common trend that the archetypal profiles are more *extreme* than the cluster centers in the sense that the values of each entry is closer to zero or one. This is very much in line with the intuition that

1. http://users.ics.aalto.fi/jhollmen/BernoulliMix/

archetypal profiles exist on the boundary of the set of observation. For binary observations, such representation has the benefit of being more easily interpretable in terms of presence and absence of attributes. For example, consider the first and fourth archetypes in DNA (Figure 8: top-left) and the corresponding cluster centers.

Second, the archetypes are further away from each other than the cluster centers, and thus, they are more easily distinguishable from each other. For example, consider the first and third archetypes of SPECT (Figure 8: top-right) and corresponding centers. This is more visible in RANDOM (Figure 8: bottom-right) which does not have any cluster structure by design.

Third, when the observations can be clustered easily, the solutions found by both methods are almost the same, e.g., see DNA, and SPECT (Figure 8: top), whereas they become very different when clear clusters do not exist, e.g., see CONGRESS and RANDOM (Figure 8: bottom). Indeed if we assign observations to archetypal profiles by the maximum $\mathbf{H}$ value, then the Rand index between this assignment and that of the clustering solution for the four datasets are (Rand index by chance within brackets), DNA: 0.90 (0.58), SPECT: 0.80 (0.62), CONGRESS: 0.77 (0.58), and RANDOM 0.65 (0.6). Rand index [15] is defined as the proportion of observation pairs that either fall in the same cluster or in different clusters in both solutions together, i.e., given

$$\alpha = \left|\left\{(x_i, x_j) : i \neq j, x_i, x_j \in C_k^1, x_i, x_j \in C_l^2\right\}\right|$$
$$\beta = \left|\left\{(x_i, x_j) : x_i \in C_k^1, x_j \in C_l^1, x_i \in C_m^2, x_j \in C_n^2\right\}\right|,$$

the Rand index between $\mathcal{C}^1 = \{C_k^1\}$ and $\mathcal{C}^2 = \{C_l^2\}$ is defined as $\mathrm{RandIndex}(\mathcal{C}_1, \mathcal{C}_2) = (\alpha + \beta)/n(n-1)/2$. A higher Rand index implies better conformity between two assignments.

4 Experiments

In our experiments, we run the VB approach with $K = 20$ archetypes and 100 trials, and present the solution with the maximum VB lower bound. Figure 9 shows the VB lower bound versus the number of active archetypes for each of the datasets, along with the best solution. To interpret the prototypes qualitatively, we *binarize* them, i.e., given a threshold τ, for each view, say marital status, we present the categories, e.g., single, married, divorced whose value in the archetypal profile is greater than τ (it is easily seen that if $\tau > 0.5$ then there is at most one active category for each view). However, instead of investigating only a specific threshold, we explore results for a few of them: a relaxed threshold reveals commonalities among prototypes whereas a strict threshold reveals prototype specific attributes. Albeit, it is easily seen that the categories present in the stricter threshold are also present in the relaxed threshold.

4.1 Austrian guest survey

The analysis of binary survey data is important in social sciences. Binary questions in comparison to a corresponding multi-category format are quicker, perceived easier, equally reliable; and the managerial implications derived do not substantially differ [16]. In this example, we analyze binary survey data from the Austrian Guest Survey conducted in the winter season of 1997 (for a cluster analysis of the summer season of 1997 see [17]). The goal is to identify archetypal profiles of winter tourists. This enables, e.g., to target potential winter tourists by sending very specific advertising material. The data consists of 1571 tourists. For each tourist 45 variables are collected: Part A (25) of the variables describes whether the tourist is engaged in a certain winter activity (e.g., alpine skiing, relaxing, or shopping); Part B (6) the accommodation (e.g., hotel or private room); Part C (1) the gender; Part D (5) the company (e.g., alone or with family); and Part E (8) the source of information (e.g., from a brochure or the Internet). In the best solution five archetypes are active (see Figure 9a).

We present the prototypes obtained by archetypal analysis and Bernoulli mixture modeling in Table 1. We observe that the prototypes are very similar to each other. This is expected from the high Rand index between the assignments to prototypes for two methods: 0.82 (0.62 by chance). However, the archetypes are more extreme representation, and thus, provide additional information about the prototypes at any given *binarization threshold*. Below we provide a brief overview of the prototypes.

2 At threshold 0.9, the second prototype does not reveal any information in the clustering solution, but the archetype still indicates a 'minimal' *person who comes with partner to enjoy alpine ski, and additionally relaxes and goes to dine* (threshold 0.7).

3 The third prototype is almost the same in two cases, and points toward a person *who comes alone to just relax without any activity*.

1 The first prototype is very similar to the second prototype and indicates to *person who only enjoys alpine skiing, but does not prefer dining and relaxation*.

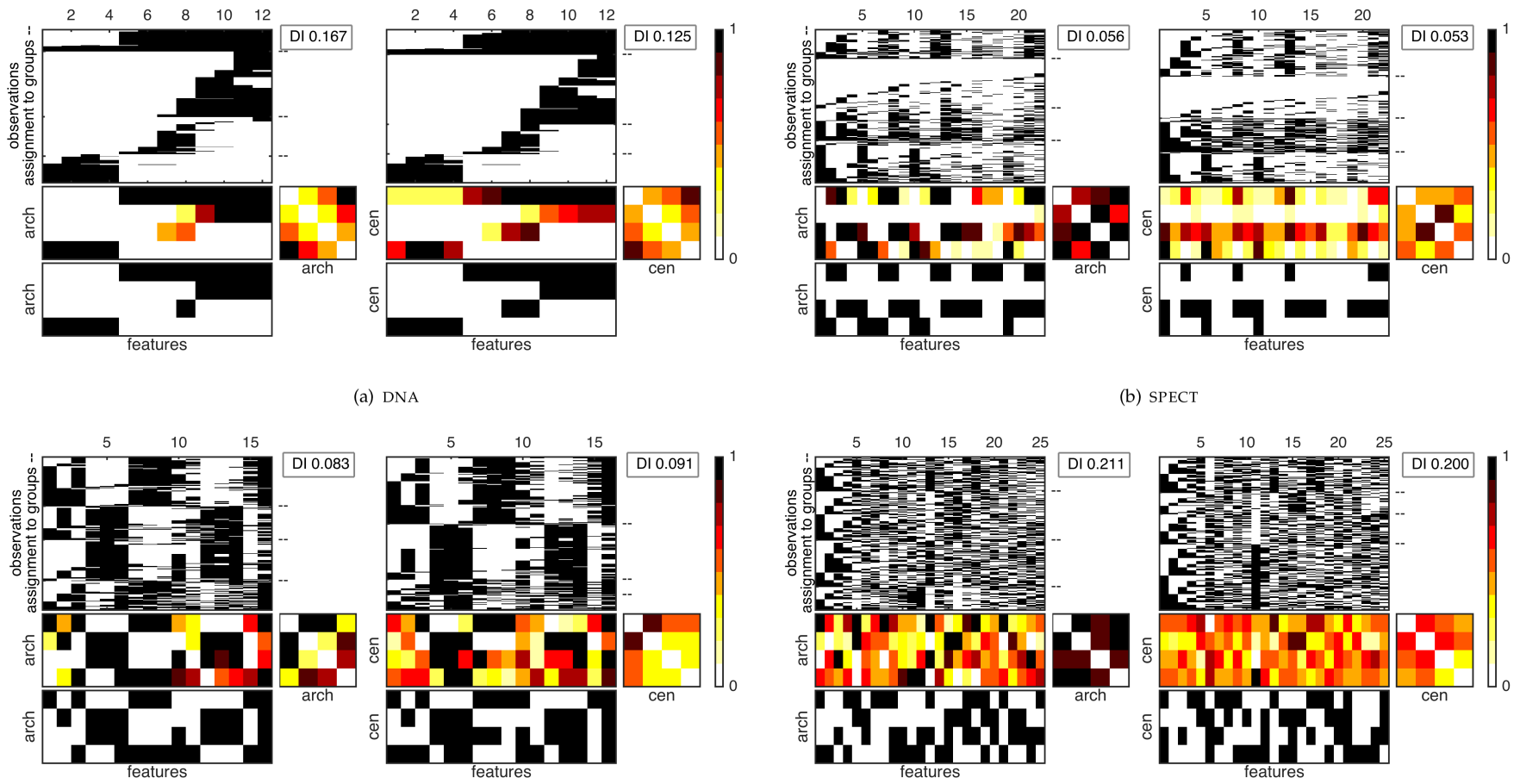
4 The fourth prototype points toward a tourist *who is not sportive and mostly comes for indoor excursion and sightseeing*.

5 The fifth prototype, very similar to the fourth prototype, however, points toward a person who is a bit *more active and enjoys other less-sportive activities such as sauna, sledge, hiking and also goes to local events*.

These archetypal profiles are very similar to the ones found in [9].

4.2 German credit dataset

In this example, we analyze a dataset in which people and their credit applications are described. Given are a set of multiple-choice attributes together with a classification into having "good" or "bad" credit risks (dataset available from [14]). The data consist of 1000 people. We use five out of 21 possible features (number of categories in bracket): credit purpose (10), employment period (5), personal status and sex (4), job situation



(a) DNA (b) SPECT

(c) CONGRESS (d) RANDOM

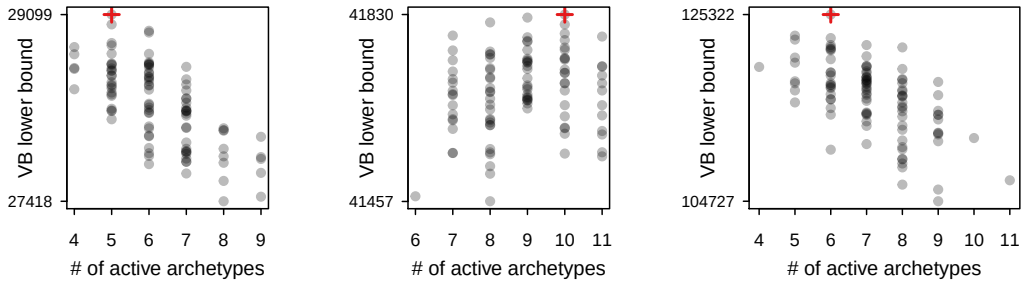
Fig. 8. Comparison of archetypal analysis and clustering of binary observations on four datasets. The datasets (a)-(d) are arranged along increasing difficulty of the task: while DNA has clear cluster structure, RANDOM does not have any. For each sub-figure the top (black and white) boxes show the assignments of observations to archetypes (left) or clusters (right). The flatter bottom (black and white) boxes show the representative observations as prototypes in the same order as the groups in top (black and white) boxes. The assignments are similar if the task is easier. The flatter boxes in between the two black and white boxes show the respective protopypes (archetypal profiles for AA and cluster centers for clustering method). It can be observed that the prototypes achieved by the two methods are more different if the task is more difficult. The square boxes on the right side of the prototype boxes show the pairwise distance between prototypes. For better visibility, the values are normalized by the largest distance achieved by both methods. It is observed that archetypes are further apart than the cluster centers. For more details see Section 3.

TABLE 1
Prototypes obtained by archetypal analysis and Bernoulli mixture modeling for Austrian guest survey (Section 4.1).

	Archetypes			Centers		
5	alpine ski sledge pool/sauna hiking walk ind. excursion relax shopping sightseeing local event hotel partner	sledge pool/sauna walk relax local event hotel partner	walk relax hotel partner	hotel partner	walk relax hotel partner	pool/sauna walk relax shopping hotel partner
4	walk ind. excursion relax shopping sightseeing hotel partner	walk ind. excursion relax sightseeing hotel partner	walk ind. excursion sightseeing partner	walk sightseeing	walk ind. excursion relax sightseeing partner	walk ind. excursion relax sightseeing partner
1	alpine ski hotel partner none	alpine ski hotel partner	alpine ski hotel	alpine ski hotel	alpine ski hotel partner	alpine ski hotel partner
3	relax hotel alone friends	relax alone	alone	alone	alone	relax alone
2	alpine ski relax dinner partner	alpine ski dinner partner	alpine ski partner		alpine ski dinner partner	alpine ski relax dinner partner
	0.7	0.8	0.9	0.9	0.8	0.7
		$\longrightarrow$	Threshold		$\longleftarrow$	

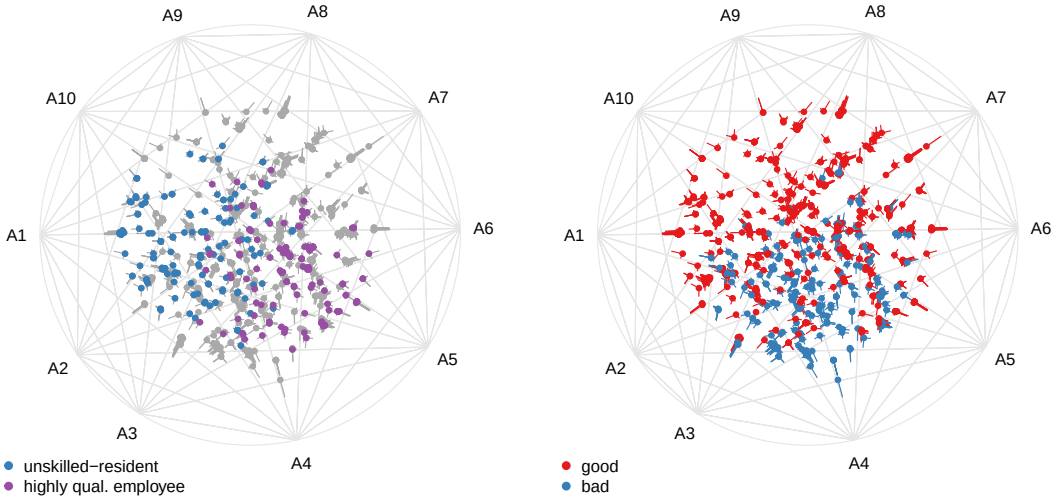
TABLE 2
Prototypes obtained by archetypal analysis and latent Dirichlet allocation for German credit dataset (Section 4.2). The table shows the three prototypes explained in more detail in the text; the full table with all ten prototypes can be found online at `http://aalab.github.io`.

	Archetypes			Topics		
7	domestic appliances $\ldots \geq 7$ years male, single skilled employee good **42**	domestic appliances $\ldots \geq 7$ years male, single skilled employee good **42**	domestic appliances $\ldots \geq 7$ years male, single skilled employee good **42**	furniture $\ldots \geq 7$ years female unemployed bad **0**	furniture $\ldots \geq 7$ years female unemployed bad **0**	furniture $\ldots \geq 7$ years female unemployed bad **0**
3	$*$ $1 \leq \ldots < 4$ years male, single skilled employee $*$ **112**	$*$ $1 \leq \ldots < 4$ years male, single skilled employee $*$ **112**	$*$ $1 \leq \ldots < 4$ years male, single skilled employee $*$ **112**	$*$ $\ldots \geq 7$ years male, separated unskilled-resident bad **1**	$*$ $\ldots \geq 7$ years male, separated unskilled-resident bad **1**	radio/televison $\ldots \geq 7$ years male, separated unskilled-resident bad **0**
4	car (new) $\ldots < 1$ year female skilled employee bad **6**	car (new) $\ldots < 1$ year female skilled employee bad **6**	car (new) $\ldots < 1$ year female skilled employee bad **6**	$*$ $*$ $*$ $*$ bad **300**	$*$ $*$ female $*$ bad **109**	$*$ $*$ female unemployed bad **5**
	0.7	0.8	0.9	0.9	0.8	0.7
		$\longrightarrow$	Threshold		$\longleftarrow$	



(a) Austrian guest survey (b) German credit dataset (c) SUN attribute dataset

Fig. 9. The VB lower bound versus the number of active archetypes for the experiments (Section 4). The best solution out of 100 trials is marked by a red plus.



(a) Job (b) Credit risk

Fig. 10. Simplex visualization for the German credit dataset. The individual factors $\mathbf{h}_j$ are projected into the circle spanned by the K archetypes (see [9] for a detailed explanation of the simplex visualization). Figure (a) highlights the two options "unskilled-resident" (blue) and "highly qualified employee" (purple) of the *Job* attributes. We can see that "unskilled-resident" observations are mainly located towards A1; and "highly qualified employee" observations are mainly located towards A4. Figure (b) highlights the two options "good" (red) and "bad" (blue) for the *Credit risk* attribute. We can see that "good" observations are basically located everywhere, but "bad" observations are located towards A3 and A4.

(4), and credit risk (2). The goal in this example is to identify archetypal profiles of credit applications. In the best solution 10 archetypes are active (see Figure 9b). We present the prototypes achieved by archetype analysis and polylingual latent Dirichlet analysis[2] in Table 2[3]. For the sake of visualization, if no categories exceed the *binarization threshold* for a feature, we represent it as a 'wildcard' ($*$).

The objective of archetypal analysis is to find patterns that are realistic, compared to topics that can be abstract concepts. To observe the 'reality' of a prototype we observe if there are actual observations that match the active categories, i.e., we compute the number of observations where the active categories are present, and the wildcard categories can be anything. These numbers are reported in bold in the table. The Rand index between two solutions are 0.79 (0.78 by chance).

Below we present a brief overview of the prototypes,

7 the seventh topic points to a *bad credit risk situation where a person, female and unemployed, but with more than seven years of experience, is buying furniture.* However, such observation is not present in the dataset. On the contrary, there are 42 observations present for the corresponding archetypal profile, a *good credit risk situation where a person, male and skilled with more than 7 years for experience, is buying domestic appliances.*

4 the fourth archetype points to a *bad credit risk situation where a person, female and skilled, but with only less than a year of experience is buying a new car*. There are

2. https://bitbucket.org/trickytoforget/polylda
3. Complete table available online at `http://aalab.github.io`

TABLE 3
Prototypes obtained by archetypal analysis and Bernoulli mixture modeling for SUN attribute dataset (Section 4.3).

	Archetypes			Centers		
4	sailing/ boating swimming natural light natural open area far-away horizon	sailing/ boating open area	sailing/ boating open area		open area	natural light open area
1	competing sports exercise natural light open area	competing sports exercise	competing sports exercise		natural light open area	natural light open area
5	enclosed area no horizon	enclosed area no horizon	enclosed area		enclosed area no horizon	enclosed area no horizon
2	no horizon					man-made
3	natural light man-made open area	natural light man-made open area	natural light man-made		natural light man-made	natural light man-made open area
6	trees vegetation foliage leaves natural light open area	vegetation foliage natural light open area			vegetation foliage natural light	trees vegetation foliage natural light open area
	0.7	0.8	0.9	0.9	0.8	0.7
		⟶	Threshold		⟵	

6 such instances. This is the only archetype where a credit risk is found to be bad.

3 the third topic points to a *bad credit risk situation where a person, male and unskilled but with more than 7 years of experience is buying a television*. However, there are no such observation in the dataset.

In Figure 10 we use simplex visualizations to interpret certain aspects of the computed archetypes solution to a greater extent. The stochastic nature of $\mathbf{h}_n$ implies that $\mathbf{Z}\mathbf{h}_n$ exists on a standard $(K-1)$-simplex with the K archetypes $\mathbf{Z}$ as the corners, and $\mathbf{h}_n$ are the coordinates with respect to these corners. A standard simplex can be projected to two dimensions via a skew orthogonal projection, where all the vertices of the simplex are shown on a circle connected by edges. The individual factors $\mathbf{h}_n$ can be then projected into this circle. We refer to [9] for a detailed explanation of the simplex visualizations. Figure 10a shows a clear discrimination between the options "unskilled-resident" and "highly qualified employee" of the Jobs attribute. Figure 10b, on the other hand, shows for the *Credit risk* attribute, no clear pattern in terms of "good" observations are visible. For "bad" observations, however, we can see a clear pattern towards the archetypes A3 and A4.

4.3 SUN attribute dataset

In this example, we analyze the SUN attribute image database: a subset of SUN image database where each image has been manually labeled to have an attribute or not [18]. There are 14340 images and 102 attributes. Each image has been labeled by 3 independent annotators, and thus, for each attribute there are 0 to 3 votes. We binarize each attribute value: if an attribute has at least 2 votes then we consider it to be present (1) or absent otherwise (0). Our goal in this example is to identify the archetypal image profiles, and explore their nature[4]. In the best solution 6 archetypes are active (see Figure 9c). We present the prototypes achieved by archetypal analysis and Bernoulli mixture model in Table 3.

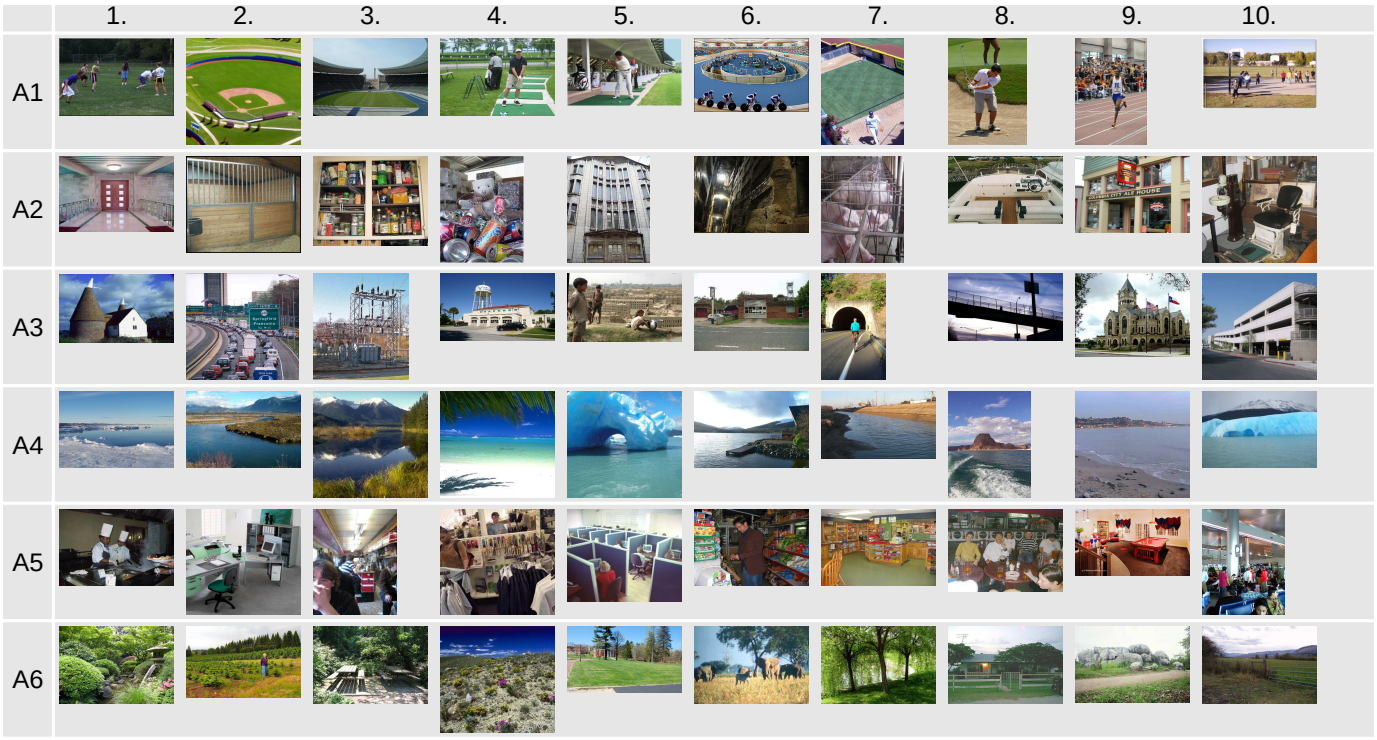
We observe that the profiles obtained by archetypal analysis are more diverse than those obtained by clustering. The Rand index between two solutions are 0.76 (0.67 by chance). Also, for a higher threshold value the clustering solutions are not present. For example,

1,4 the first and fourth archetypes (threshold 0.7) point towards an *image of a natural place (with natural light) that is open (with far away horizon) and is related to water related activities such as sailing, swimming*, and *image of an open area with natural light but related to physical activity such as sport, competition, and exercise*. Both these prototypes are missing in the clustering solution. The corresponding clustering solutions are very similar to each other, and to fifth and third prototype. A possible reason being that these attributes are present in most images, and thus clustering solution tries to model those.

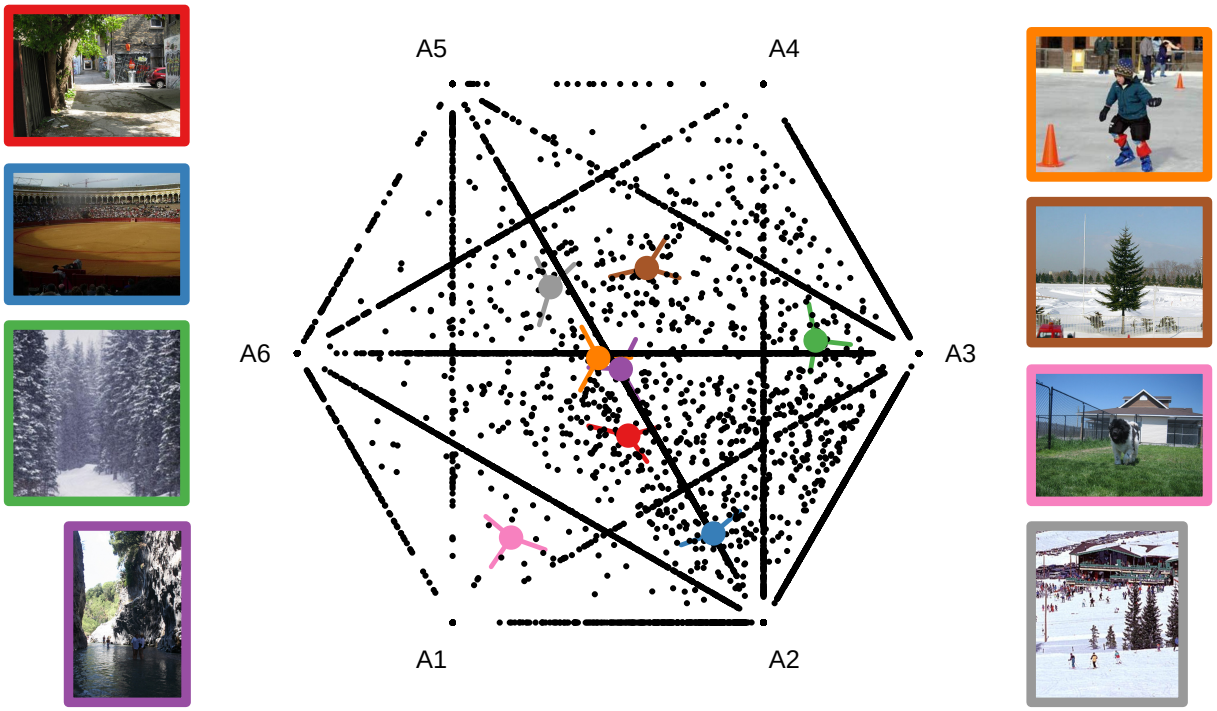
3,5,6 the sixth, third, and fifth prototypes (threshold 0.7) match quite well between two solutions, and points toward an *image of an open area with natural light and greenery*, *image with natural light and open area but having man made objects*, and *image with enclosed area and without a horizon*.

2 the second prototype (threshold > 0.7) is also similar in both solutions and points toward an *abstract image with no particular attributes*.

4. Note that the corresponding $\mathbf{h}_n$ can be used in image retrieval.



(a)



(b)

Fig. 11. Visualization of the archetypal analysis solution for the SUN attribute dataset. (a) The top ten generating images for each one of the six archetypes. (b) The simplex visualization with eight images highlighted that are composed of three archetypes. For example: the blue image is composed of A1 (competing, sports exercise), A2 (no horizon), and A3 (natural light, man-made); the red image is composed of A2 (no horizon), A3 (natural light, man-made), and A6 (trees, vegetation, foliage); the orange image is composed of A1 (competing, sports, exercise), A3 (natural light, man-made), and A5 (enclosed area, no horizon).

We present the generating (observations with $w_{ik} > \epsilon$) images for each (k-th) archetype in Figure 11 and also highlight eight images where each image is composed by three archetypes. We observe that both the generating images and the composed images are rather intuitive. For example, the image of a bullring (blue) (for bullfighting, not a sport) is composed of A1 (sports activity), A2 (abstract), and A3 (man-made), whereas the generating images of A1 (sports activity) are all related to popular sports, and usually with people actively participating.

5 Conclusion

In this paper, we have introduced probabilistic archetypal analysis for nominal observations, e.g., multiple-option questionnaires. The core of this extension is to construct a generative model which ultimately treats each feature as an independent archetypal analysis problem with multinomial observation model. This construction allows us to derive efficient update rules using principle of variational Bayes'. Additionally, it provides a principled approach for selecting appropriate number of archetypes by utilizing suitable prior information over the coefficient vectors, which so far has been done by the elbow criterion, a heuristic approach. Together these extensions expand the applicability of archetypal analysis to a wider range of practical problems, e.g., marketing research where multiple option questionnaires are popularly used to understand customer behavior. We have demonstrated the effectiveness of the proposed approach over related methods such as clustering and topic modeling on three real world questionnaire datasets.

Although the proposed approach has been formulated in the context of nominal variables, it is actually a generic formulation since one can always view a d-dimensional archetypal observation problem as d independent single dimensional problem, and each subproblem can originate from a separate exponential family distribution. This allows extending archetypal analysis to tackle mixed data types. An issue with the proposed approach is the computational complexity. It increases quadratically with the number of samples, and linearly in number of features. Recently, faster methods for archetypal analysis is being explored actively [19], and this remains to be explored for the current extension.

References

[1] B. H. P. Chan, D. A. Mitchell, and L. E. Cram, "Archetypal analysis of galaxy spectra," *Monthly Notices of the Royal Astronomical Society*, vol. 338, no. 3, pp. 790–795, 2003.

[2] Y. Xiong, W. Liu, D. Zhao, and X. Tang, "Face recognition via archetype hull ranking," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 585–592.

[3] M. J. A. Eugster, "Performance profiles based on archetypal athletes," *International Journal of Performance Analysis in Sport*, vol. 12, no. 1, pp. 166–187, 2012.

[4] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[5] C. Bauckhage and C. Thurau, "Making archetypal analysis practical," in *Pattern Recognition*, ser. Lecture Notes in Computer Science, J. Denzler, G. Notni, and H. Se, Eds. Springer Berlin Heidelberg, 2009, vol. 5748, pp. 272–281.

[6] C. Thurau, K. Kersting, and C. Bauckhage, "Yes we can: Simplex volume maximization for descriptive web-scale matrix factorization," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, p. 17851788.

[7] M. J. A. Eugster and F. Leisch, "Weighted and robust archetypal analysis," *Computational Statistics and Data Analysis*, vol. 55, no. 3, pp. 1215–1225, 2011.

[8] M. Mørup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," *Neurocomputing*, vol. 80, pp. 54–63, 2012.

[9] S. Seth and M. J. A. Eugster, "Probabilistic archetypal analysis," *Machine Learning*, pp. 1–29, 2015, available as "Online First". [Online]. Available: http://dx.doi.org/10.1007/s10994-015-5498-8

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[11] T. Hofmann, "Probabilistic latent semantic analysis," *arXiv:1301.6705*, 2013. [Online]. Available: http://arxiv.org/abs/1301.6705

[12] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 880–889. [Online]. Available: http://dl.acm.org/citation.cfm?id=1699571.1699627

[13] J. Hollmén and J. Tikka, "Compact and understandable descriptions of mixtures of bernoulli distributions," in *Proceedings of the 7th International Conference on Intelligent Data Analysis*, ser. IDA'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 1–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=1771622.1771624

[14] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[15] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. pp. 846–850, 1971. [Online]. Available: http://www.jstor.org/stable/2284239

[16] S. Dolnicar, B. Grün, and F. Leisch, "Quick, simple and reliable: Forced binary survey questions," *International Journal of Market Research*, vol. 53, no. 2, pp. 231–252, 2011.

[17] S. Dolnicar and F. Leisch, "Segmenting markets by bagged clustering," *Australasian Marketing Journal*, vol. 12, no. 1, pp. 51–65, 2004.

[18] G. Patterson and J. Hays, "SUN attribute database: Discovering, annotating, and recognizing scene attributes," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2012, pp. 2751–2758.

[19] Y. Chen, J. Mairal, and Z. Harchaoui, "Fast and robust archetypal analysis for representation learning," *CoRR*, vol. abs/1405.6472, 2014.

Sohan Seth received his PhD in electrical and computer engineering from the University of Florida, Gainesville, in 2011. He was a postdoctoral researcher in the Aalto University, Finland till 2014. He is currently a research associate in the University of Edinburgh, UK. His research interests include machine learning and computational biology.

Manuel J. A. Eugster received his PhD in statistics from the Ludwig-Maximilians-Universität München, Germany, in 2011. He is currently a postdoctoral researcher in the Aalto University, Finland. His research interests include machine learning in the intersection of information retrieval and brain-computer interfaces.